

Web3 Meets AI Marketplace: Exploring Opportunities, Analyzing Challenges, and Suggesting Solutions

MintAI Protocol Labs^{a,1}

^aIntellicain Solutions Kft

Abstract—Web3 and AI have been among the most discussed fields over the recent years, with substantial hype surrounding each field's potential to transform the world as we know it. However, as the hype settles, it's evident that neither AI nor Web3 can address all challenges independently. Consequently, the intersection of AI and Web3 is gaining increased attention, emerging as a new field with the potential to address the limitations of each. In this article, we will focus on the integration of web3 and the AI marketplace, where AI services and products can be provided in a decentralized manner (DeAI). A comprehensive review is provided by summarizing the opportunities and challenges on this topic. Additionally, we offer analyses and solutions to address these challenges. We've developed a framework that lets users pay with any kind of cryptocurrency to get AI services. Additionally, they can also enjoy AI services for free on our platform by simply locking up their assets temporarily in the protocol. This unique approach is a first in the industry. Before this, offering free AI services in the web3 community wasn't possible. Our solution opens up exciting opportunities for the AI marketplace in the web3 space to grow and be widely adopted.

Contents

1	Introduction	1
1.1	Web3 at a glance	1
1.2	AI's explosive growth	1
2	Opportunities and Challenges	2
2.1	Why is web3 more accessible?	2
2.2	Web3 infrastructure as an advantage	2
2.3	Challenges of Merging AI and Web3 Infrastructure	3
3	Analyzing Solutions	4
3.1	L1, L2 or L1-L2 architecture?	4
3.2	Economic model analysis: charged vs. uncharged approaches	4
	<i>Uncharged approaches for decentralized AI services</i>	
3.2.1.1	The system's robustness in general	5
3.2.1.2	The miner's perspectives	5
3.2.1.3	Rewards for miners	5
3.2.1.4	The client's perspectives	6
3.2.1.5	The coordinator's aspect and system securities	6
3.2.1.6	Securities	6
3.2.1.7	Economics	6
	<i>Charged approaches for decentralized AI services</i>	
3.3	Protocol implementations: network participants and process flow	7
3.3.0.1	The global ledger, coordinator nodes and peer to peer connections	8
3.3.0.2	Client cycle	8
3.3.0.3	Mining Cycle (for service providers)	9
3.4	Discussions	9
	<i>Capability and scalability • Security • Advantages</i>	
4	Conclusions and Future Works	10
5	Contact us	10
6	Supporting References	10

1. Introduction

Artificial Intelligence (AI) and Web3, as two big players in the world of tech, have caught everyone's attention in recent years. Think of AI as the capability of computers to mimic human intelligence, helping us do things like assisting programming, driving cars, or even diagnosing illnesses. On the other hand, Web3 is like a new version of

the internet, where instead of a few big central entities holding all the power and data, it's distributed, making everything more transparent and fair.

1.1. Web3 at a glance

Web3, often termed as the "new internet", is the next phase in the progression of the World Wide Web (WWW). If web1.0 was about static web pages and read-only content, and web2.0 brought interactivity, social media, and user-generated content, then web3 is about decentralized and trustless protocols and technologies. It moves away from centralized control and ownership, as seen with big tech companies today.

At its core, web3 is primarily built on decentralized blockchain technology. It emphasizes user control over personal data, trustless interactions (meaning you don't need third-parties to trust each other), and direct peer-to-peer exchanges of value. In some situations, web3 can perform tasks more effectively than traditional web applications, and in certain cases, it can achieve what traditional platforms cannot. Below, we summarize some of the hottest sectors in web3:

- 1. Decentralized Marketplaces:** Peer-to-peer marketplaces where users can transact directly without middlemen. This applies to both goods and services. Decentralized finance (DeFi) stands out as one of the most significant sectors in decentralized marketplaces which aims to recreate traditional financial systems, such as loans, savings, insurance, and more, in a decentralized manner using smart contracts on blockchains. Unlike traditional finance, DeFi operates in a fixed and transparent manner, and there is no room for hidden activities behind the scenes in such financial products.
- 2. Borderless Transactions:** Traditional financial systems often impose high fees and delays on international transfers. Cryptocurrencies, such as Ripple [7], allow for almost instantaneous global transactions with minimal fees amounting to mere cents.
- 3. Digital Authenticity:** Traditional digital files can be copied endlessly, making it hard to identify the "original." Non-Fungible Tokens (NFTs), on the other hand, provide a unique stamp of authenticity that can't be duplicated. Every sale or transfer is transparently recorded, ensuring true ownership and history. This means artists and creators can sell their work digitally, knowing there's a verifiable "original" out there. NFTs have gained massive traction in art, collectibles, and even real estate in the virtual space.
- 4. Decentralized Decision Makings:** Traditional organizations have a hierarchical structure where decisions often come from the top. DAOs (Decentralized Autonomous Organizations) operate on consensus mechanisms, allowing all members to have a say. Without a central authority, decisions can be made transparently and collectively, ensuring every stakeholder's voice is heard and reducing the risks of centralized corruption or biases.
- 5. Decentralized Web Infrastructure:** This includes decentralized file storage solutions with platforms like Filecoin [10] and IPFS (InterPlanetary File System) [6], GPU service systems [41], blockchain oracles [21] and more, which ensure that the foundational aspects of the internet are distributed and not controlled by any single entity.

1.2. AI's explosive growth

Previously, notable AI milestones, such as AlphaGo [8], which defeated a world champion Go player in 2016, were celebrated but

64 quickly faded from the public's consciousness. Later on, the global
 65 financial market witnessed the explosive growth of generative AI
 66 (gen AI) tools in 2022, capable of producing various types of content,
 67 including text, imagery, audio, and synthetic data. Among these tools,
 68 OpenAI's GPT implementation stands out as the AI-powered chatbot
 69 that took the world by storm. Driven by incredible market reactions,
 70 it is estimated that ChatGPT reached 100 million monthly active users
 71 in January 2023 [36]. In comparison, TikTok took nine months to
 72 achieve 100 million users, while Instagram took 2.5 years. Experi-
 73 ments show that ChatGPT significantly increased productivity, with
 74 the average time taken reduced by 40% and output quality improving
 75 by 18% [40].

76 Although generative AI caught the public's attention in 2023, AI is
 77 generally considered to include other main sectors such as supervised
 78 machine learning, unsupervised machine learning, and reinforce-
 79 ment learning. The marketplace of AI as a whole is expanding rapidly.
 80 The global AI solution market is projected to attain 301.2 billion USD
 81 by 2028, with a compound annual growth rate (CAGR) of 29.4%.
 82 Specifically, the unsupervised machine learning sector is forecasted
 83 to reach 15.6 billion USD by 2028, expanding at a CAGR of 25.1%.
 84 Furthermore, AI solutions deployed in public cloud environments
 85 are expected to triple the figures of private cloud implementations
 86 during the same period [34]. Regarding regions, North America gen-
 87 erated more than 36.84% of the market share in 2022. The Asia Pacific
 88 market is expected to expand at the highest CAGR of 20.3% from 2023
 89 to 2032 [32].

90 While AI and Web3 have their distinct advantages both in utilities
 91 and marketplace, a growing area of interest lies in their combined
 92 potential. The idea is simple: what if we could merge the decentral-
 93 ized approach of Web3 with the capabilities of AI? This combination
 94 could lead to more powerful AI systems that are also more accessible
 95 to everyone. Consider the possibilities, such as using Web3's decen-
 96 tralized structure to train AI models, or making AI tools available to
 97 a wider audience through Web3 platforms.

98 In this article, our focus is on understanding how AI and Web3
 99 can be merged, examining the potential opportunities and addressing
 100 the challenges. We will dive into topics currently of great interest
 101 in this field. Our aim is to provide a clear and informed perspec-
 102 tive on the research intersections between these two transformative
 103 technologies.

104 **2. Opportunities and Challenges**

105 In this paper, we are particularly interested in the integration of the
 106 web3 infrastructure with the AI marketplace, an area where web3
 107 enhances AI's product performance in the marketplace. While there
 108 are various ways that AI and web3 can complement each other, such
 109 as using AI to produce NFTs or employing AI as an autopilot for code
 110 writing, our focus is on the opportunities in the current commercial
 111 domain. However, we approach this with a broader vision, digging
 112 into the potential of both AI and web3. In following subsections, we
 113 will explore the opportunities and challenges of such an integration.

114 **2.1. Why is web3 more accessible?**

115 According to the statistics of the world bank in 2021 [23], credit card
 116 ownership seemed to align with a nation's development. Canada
 117 led with 82.7%, followed closely by developed countries like Israel,
 118 Iceland, and Japan. The USA stood at 66.7%. Many European nations
 119 reported over 50% ownership. In contrast, many African and South
 120 Asian countries, such as Nigeria and Pakistan, recorded less than 2%.
 121 The trend suggests that developed countries have a higher percentage
 122 of credit card holders compared to less economically advanced na-
 123 tions. As a result, people in generally less-developed countries have
 124 limited access to paid AI services due to a lack of payment methods.
 125 Furthermore, teenagers under the age of 15 typically have limited
 126 access to credit cards. Consequently, they have restricted access to
 127 AI services and support. If they wish to use these services, they often

have to rely on their parents' cards, creating transactional friction
 128 and making AI less accessible to the general public.

129 Web3 and crypto, however, offer a much simpler payment pro-
 130 cess. Taking Ethereum as an example, since anyone can create an
 131 Ethereum wallet easily by setting up everything on their mobile with
 132 a few clicks, it's more user-friendly than traditional banking and
 133 centralized payment methods. And having an Ethereum wallet and
 134 ether means you have access to the web3 world, so almost everyone
 135 can access web3 if they buy any type of crypto. Moreover, people are
 136 starting to accept crypto as salaries since they don't have to use an
 137 international bank to receive it [22]. Famous comments, such as one
 138 from Vitalik in 2020, mention workers in Africa accepting ETH as
 139 payment. A broader and simpler access through web3 could make
 140 AI services more available and exciting as an industry.

141 **2.2. Web3 infrastructure as an advantage**

142 Any public chain requires a consensus mechanism to update the
 143 global states in a distributed network system, with Proof of Work
 144 (PoW) being the most commonly used [14], [20]. The consensus
 145 mechanism is often referred to as crypto mining, which involves
 146 creating and adding new blocks to a blockchain network using var-
 147 ious consensus methods based on different resources (like mining
 148 rigs, staked tokens, etc.). In the PoW consensus mechanism, miners
 149 compete to produce the next valid block by being the first to solve a
 150 cryptographic puzzle, thereby earning a reward for their efforts. In
 151 the marketplace, as a public chain gains popularity, its crypto miners
 152 receive increased rewards. This attracts more miners, or in other
 153 words, more computing power, to join the chain. Consequently, the
 154 total hash power of the chain continues to rise over time. Notably,
 155 prominent blockchain projects in the crypto industry, such as Bitcoin
 156 (BTC) and Ethereum (ETH), have used the PoW consensus mecha-
 157 nism for years. According to Bitcoin energy consumption analysis
 158 [24], [31], the annual electricity consumption of Bitcoin mining sur-
 159 passed that of the United Arab Emirates (119.45 TWh) in 2021 and
 160 Sweden (131.79 TWh) in 2022. Most of this energy is dedicated to
 161 solving cryptographic puzzles.

162 While this process achieves trustless consensus, it doesn't offer
 163 practical benefits beyond producing block hashes that, in Bitcoin's
 164 case, have a certain number of zeros at the beginning [5]. Conse-
 165 quently, the absence of a theoretical limit on energy consumption
 166 for the PoW mechanism has raised global concerns. This led to the
 167 exploration of alternative consensus mechanisms, like Proof of Stake
 168 (PoS), and changes in institutional policies. For example, in 2021,
 169 Tesla announced it would no longer accept BTC due to climate con-
 170 cerns [27]. In 2022, Ethereum transitioned from the energy-intensive
 171 Proof of Work (PoW) mechanism to the more efficient Proof of Stake
 172 (PoS) in response to environmental concerns. This shift resulted in
 173 a significant reduction in energy demand, with decreases ranging
 174 from 99.84% to 99.9996% [35]. This reduction in Ethereum's energy
 175 consumption is comparable to the electricity needs of countries like
 176 Ireland or even Austria, marking a notable step towards environmen-
 177 tal sustainability. However, this change also left a large amount of
 178 unused hashrate, equivalent to 1,126,674 GH/s [37], without a specific
 179 use. The advancement of computing resources in crypto mining isn't
 180 the only type of resource in web3. Linear or exponential growth in
 181 the infrastructure supporting specific consensus algorithms has been
 182 widespread in most mainstream web3 utility projects. For instance, in
 183 cloud storage, the capacity of decentralized storage has surged from
 184 just a few million TB to nearly a hundred EiB¹, as per reports and
 185 statistics [29], [38]. Furthermore, the cost of decentralized storage is
 186 on average \$0.19 per month, which is much cheaper than centralized
 187 solutions such as Dropbox.

188 Meanwhile, as artificial intelligence (AI) becomes integrated into
 189 various sectors of the economy, there's a rapidly growing demand for
 190 computational resources to power this machine intelligence. Train-

1¹ EiB = 1,152,921,504.6068 GB

192 ing a model like ChatGPT can cost over \$5 million, and the initial
 193 operation of the ChatGPT demo ran OpenAI an approximate \$100,000
 194 daily, even before its current usage surged [33]. Midjourney, a service
 195 that provides high-quality images, operates with more than 9,000
 196 GPU cards, contributing to its operational costs. Given the vast num-
 197 ber of neural parameters and extensive GPU hours involved, the high
 198 computational demands of model optimization pose significant chal-
 199 lenges for academic researchers and small-scale enterprises. This
 200 limits the broader adoption and use of artificial intelligence technolo-
 201 gies.

202 It is, therefore, unsurprising that an increasing number of crypto
 203 miners are exploring ways to use their existing computational in-
 204 frastructures to advance AI. They are redirecting computational re-
 205 sources, which were previously focused on mining, toward machine
 206 learning and other high-performance computing (HPC) applications,
 207 such as the Internet of Things (IoT) and data services [15], [18]. An-
 208 other example is provided by Hive Blockchain, which is shifting its
 209 long-term HPC strategy from Ethereum mining to applications like
 210 artificial intelligence, rendering, and video transcoding, contributing
 211 to their total annual revenue generation of approximately \$102 mil-
 212 lion. Miners can also opt to employ these resources for processes on
 213 decentralized blockchain networks.

214 2.3. Challenges of Merging AI and Web3 Infrastructure

215 While there are significant opportunities in both marketplaces, we
 216 have identified major challenges that prevent the development of
 217 standout applications. We will analyze these difficulties from both
 218 market and technical perspectives to better inform potential solutions
 219 for integrating web3 with the AI marketplace.

220 1. **Blockchain resources are inherently costly:** When it comes
 221 to the blockchain consensus infrastructure, resources are, by de-
 222 sign, typically expensive. The FLP impossibility theorem states
 223 that in an asynchronous distributed system, where at least one
 224 process can fail, it's impossible to design a consensus algorithm
 225 that simultaneously guarantees both safety and liveness [1]. This
 226 is a primary reason most blockchain systems adopt synchronous
 227 or partially synchronous² consensus mechanisms such as bitcoin
 228 or ethereum. However, such systems often have substantial stor-
 229 age and bandwidth costs, especially since they store n replicas of
 230 the global states. It's therefore essential for the protocol to main-
 231 tain only the necessary states, minimizing storage requirements.
 232 Given the rapid development of the AI marketplace, embedding
 233 the entire system into a layer-1 (L1) blockchain solution³
 234 might not be the most efficient strategy [17], [25]. Such systems
 235 generally uphold a consistent block production rate⁴, thereby
 236 ensuring a consistent transaction throughput capacity. However,
 237 in the case of decentralized AI marketplace, the workload can
 238 vary dynamically based on market supply and demand. There
 239 might be instances where the system witnesses inactivity due to
 240 an absence of incoming training tasks, resulting in most nodes
 241 becoming stale without a continuous reward stream. In this
 242 setup, the primary goal is to orchestrate AI market activities in
 243 the network, with transaction validation serving as a secondary
 244 role. A well-constructed framework should:

- 245 • address these aspects by dynamically adjusting system
 246 workload based on the influx of jobs and tasks
- 247 • enable seamless system upgrades over time

²There's an assumption that a time upper bound exists for message delivery and block production; however, this bound may be unknown or subject to change during system upgrades.

³A Layer-1 (L1) blockchain represents the fundamental tier of a blockchain network, comprising the base protocol that oversees the consensus mechanism, transaction processing, and data storage. This layer delivers the core functionality of the blockchain system and supplies the infrastructure for crafting additional layers or applications atop it.

⁴The block production rate in a blockchain denotes the rate at which new blocks are generated and appended to the blockchain. For instance, the TRON (TRX) network boasts a rapid block production rate, with a fresh block produced every 3 seconds.

- 248 • ensure the ease of use and security for users' assets

249 Unfortunately, such a system is currently lacking in the industry.

250 2. **Payment frictions in AI service subscriptions:** Even with
 251 the assumption that cryptocurrencies offer easier access and
 252 operation, the business revenue model for various AI services
 253 remains a challenge. While customers are willing to pay for
 254 specific tasks, they resist being charged repeatedly when switch-
 255 ing between services—a common occurrence in traditional AI
 256 businesses. For instance, if you purchase a ChatGPT premium
 257 for access to GPT-4 and additional features, you'd still have to
 258 pay separately for a Midjourney premium should you need its
 259 services. Consequently, customers wanting to use a broad array
 260 of AI services could face hundreds of dollars in monthly sub-
 261 scriptio fees. Even within the same company or network, users
 262 don't want to be charged each time they order tasks, as seen
 263 with the GPU tasks pricing model in the render network [41].
 264 Exploring how web3 solutions can enhance the user experience
 265 regarding subscription practices is of significant interest.

266 3. **Integrating multiple parties:** In the traditional AI business
 267 model, there is a direct value exchange between two main par-
 268 ties: the customers and the service providers. Similarly, in most
 269 blockchain models, there are only two primary participants: the
 270 crypto users who send the transactions and the crypto miners
 271 who validate those transactions. As evident, both traditional AI
 272 products and web3 communities involve only two major par-
 273 ties. While web3 infrastructure has the potential to broaden the
 274 accessibility of AI and offer better market rates, its integration
 275 introduces additional participants into the network, thereby in-
 276 creasing complexity. In general, there are at least three parties
 277 involved: the customer, the miner providing computing power
 278 and storage, and the product designers who contribute the foun-
 279 dational building blocks for various AI services. Developing
 280 a system framework and reward models that benefit all three
 281 major parties poses significant challenges.

282 4. **Securities:** Web3 emphasizes decentralization. However, dis-
 283 tributed systems are inherently unstable and insecure. In de-
 284 signing the system/framework we describe, we must account for
 285 a significant number of nodes being faulty or malicious up to a
 286 certain percentage. All blockchain systems employ mechanisms
 287 to prevent attackers from initiating various types of attacks. Put
 288 simply, attacking the system should be more costly financially
 289 for the attacker than the total potential reward they might gain
 290 from the attack. Consequently, different blockchain systems
 291 implement their own consensus mechanisms to prevent attackers
 292 from forging and tampering with data and states [11], [14], [20],
 293 with Proof of Work (PoW) being the most widely adopted.
 294 The consensus mechanism for integrating web3 and AI requires
 295 a novel design, as proving service provision can be quite tricky.
 296 This mechanism must account for various participant roles. Pri-
 297 marily, it needs to ensure that service providers are executing
 298 their tasks both honestly and diligently. If service providers con-
 299 duct denial of service or provide low quality service to too many
 300 customers, the system should either forfeit some of their rewards
 301 or, at a minimum, impact their reputation. This will alert future
 302 customers to be wary of these specific providers. Moreover, if a
 303 reward forfeiture or reputation system is part of the consensus
 304 mechanism, there must also be a safeguard against customers
 305 providing unjust or malicious reviews. Without a robust pro-
 306 tocol, genuine service providers could become targets of sybil
 307 attacks. Lastly, nodes responsible for maintaining global state
 308 records must be given sufficient cryptoeconomic incentives to
 309 act both honestly and diligently, given their crucial role in en-
 310 suring system security.
 311 To the best of our knowledge, such protocols addressing all the
 312 challenges mentioned above are currently lacking in the research
 313 field, and we don't see many implementations in the industry

314 field, other than fetch.AI and singularityNET [18], [42] which
 315 partially addressed the challenges. While the potential market
 316 size and areas of opportunity can be tremendous, we believe
 317 that the following challenges, as summarized from previous
 318 discussions, must be addressed to succeed in the large-scale
 319 commercialization of the AI marketplace integrated with web3
 320 infrastructure.

321 **Protocol capacity and scalability:** The system/platform
 322 should be capable of coordinating clients, miners,
 323 and AI product development, and it should empower self-
 324 governance to initiate, process, and finalize services. The
 325 volume of transactions—including client orders, reward
 326 claims, and network management—will largely be deter-
 327 mined by the customer base and the size of the AI mar-
 328 ketplace within the network. The computational power
 329 needed to maintain the system's global states should be
 330 possible to analyze theoretically. Additionally, the protocol
 331 should consider certain commercial factors, integrating
 332 "free" features that are specific to the blockchain indus-
 333 try, such as the inflation model, to enhance its appeal to
 334 potential customers.

335 **Protocol securities:** Given the nature of the AI market-
 336 place, it's impractical for a central ledger to check on every
 337 transaction, such as AI services, to ensure they're executed
 338 honestly—both in theory and practice. AI services require
 339 substantial computation, and given that many incorporate
 340 randomness and don't lead to a single definitive outcome,
 341 it's theoretically challenging for other nodes to determine
 342 if a single node is functioning accurately. Thus, a proto-
 343 col safeguarded by cryptoeconomics—where attacking the
 344 system costs more than complying with it—is preferable.
 345 To launch an attack, one would typically need more tokens
 346 than the counterparties, which can often be financially
 347 impossible. In other words, without significant potential
 348 rewards, there's little motivation to compromise the proto-
 349 col. Systems must be intricately constructed to prevent
 350 the potential rewards from being so attractive that the sys-
 351 tem's design itself becomes a target for malicious activities.
 352 When attackers recognize that their attacks will be easily
 353 corrected by the system, they have tiny incentive to pro-
 354 ceed.

3. Analyzing Solutions

355 In this section, we will propose general guidelines and possible solu-
 356 tions by analyzing the pros and cons in different architecture design
 357 and implementations. Our primary focus is on two different aspects:
 358 the technical aspects and the economic aspects. In the technical as-
 359 pects, we will focus primarily on system security and efficiency, and
 360 in the economic aspect, we will focus on customer experience and
 361 diversity in service subscriptions.

363 3.1. L1, L2 or L1-L2 architecture?

364 Blockchains based solely on L1 have their own mechanisms of pro-
 365 ducing blocks, while blockchains comprising both L1 and L2⁵ place
 366 most of the utility/core infrastructures on L1 for efficiency and move
 367 most token logistics and value storage to the L2 layer. This setup often
 368 relies on many other well-known blockchain ecosystems to serve a
 369 wider range of customers and investors. L2 solutions typically embed
 370 their program logic and database into smart contracts within main-
 371 stream ecosystems. These projects integrate their logic entirely into

372 ⁵A Layer 2 (L2) in blockchain refers to a secondary protocol or framework built on top of
 373 an existing blockchain, primarily aiming to enhance the network's scalability, efficiency,
 374 and transaction throughput. Layer 2 solutions leverage the security and decentralization
 375 of the underlying blockchain (Layer 1), while offloading a portion of the computational
 376 workload to a separate network or system. This enables faster and cheaper transactions,
 377 as well as more complex operations, without burdening the base layer. Examples of
 378 Layer 2 solutions include state channels, sidechains, and rollups.

379 smart contracts, complemented by a frontend framework connected
 380 to the backend contracts.

381 Table 1 summarized the performance matrix of different archi-
 382 tecture design. L1 ecosystems typically have their own databases
 383 and block production mechanisms. All transactions and associated
 384 state changes occur on-chain, using their local utility coins/tokens.
 385 Transaction fees ϵ can be set to very small values, as seen in the Tron
 386 network [16]. However, once initiated, such blockchains can't easily
 387 be halted, and upgrades to core functions can be challenging. Such
 388 upgrades often necessitate a hard fork by miners or validators, which
 389 requires extensive communication between various parties to adopt
 390 a new protocol at a predetermined block height [26]. In our effort
 391 to integrate web3 infrastructure with the AI marketplace, we need a
 392 setup that allows for ongoing system updates and feature additions
 393 without disrupting the network's assets or user experience. Building
 394 everything on L1 may not be the optimal solution.

395 L2 solutions, on the other hand, place all their core logic on a spe-
 396 cific public mainnet, eliminating off-chain costs. All activities occur
 397 on-chain through contract calls to the mainnet. Assisted by oracles
 398 [21], L2 ecosystems span a wide range of areas including DeFi, Gam-
 399 ing, and NFT Marketplaces. Upgrades in L2 are typically handled
 400 using the upgradable contract paradigm, where contract updates are
 401 achieved by redirecting the proxy contract pointer [28]. However,
 402 given that AI models and product upgrades cannot be fully migrated
 403 on-chain, this architecture is not suitable in the given context.

404 To effectively harness the potential of this decentralized network
 405 for web3 and AI marketplace merging, a two-layer L1-L2 architecture
 406 is introduced. The on-chain component (SC) records the value flow
 407 within the network, while the off-chain component (exec) comprises
 408 a set of protocols operating on the distributed network where utilities
 409 are executed. By seamlessly integrating the on-chain functionality
 410 with the diverse off-chain services provided, the system can achieve
 411 the robustness and upgradability that traditional Layer 1 solutions
 412 often lack. In the L1-L2 design, protocols and infrastructures mainly
 413 operate off-chain within the decentralized network, while token utili-
 414 ties like transfer and withdrawal function on Layer 2 of mainstream
 415 blockchains such as BSC or Polygon. This configuration enables the
 416 system to regularly update with new features and utilities, all while
 417 preserving the network's assets and the user experience. In the AI
 418 marketplace, the core module can be designed in L1 to ensure easy
 419 upgradability. Participants' databases can be distributed between L1
 420 and L2 by placing their assets in L2 and conducting transactions in
 421 L1, thereby increasing efficiency and reducing costs. Protocols such
 422 as Chainlink and Proof of Training (POT) [9], [39] also adopt the
 423 L1-L2 architecture.

3.2. Economic model analysis: charged vs. uncharged ap- 424 proaches

425 Most existing decentralized AI products attempt to collect micropay-
 426 ments for every user request at their own dedicated rates. As a result,
 427 these platforms require users to continually purchase cryptocurrency
 428 to pay for services and bandwidth. This suggests that the services
 429 might not be readily available for the general public to access for free
 430 via their browsers. Meanwhile, the quality of services varies widely,
 431 ranging from large-scale enterprises offering high-quality services
 432 to home computer and GPU providers renting out resources with
 433 slow internet connections. Users often remain unaware of the qual-
 434 ity of the services they are using, yet they are continuously charged.
 435 All of these create transaction frictions and prevent large scale com-
 436 mercialization and adoption of the decentralized AI apps. All of
 437 these factors create transaction frictions and prevent large-scale com-
 438 mercialization and adoption of the decentralized AI apps. In the
 439 following subsections, we will provide solutions for both charged and
 440 uncharged scenarios.

Table 1. Performance comparison of L1, L2 and L1-L2 architecture

Architecture	Mainnet	Performance				
		Transaction fees on-chain ^a	Transaction fees off-chain ^a	Supported tokens	Stablecoin integration?	Upgradability
L1	-	ϵ^b	0	local	no	difficult
L2	Ethereum	0.0004 units	0	ETH&ERC20	yes	medium
	BSC ^c	0.000075 units	0	BNB&BEP-20	yes	medium
	Tron	0.027 units	0	Tron&TRC20	yes	medium
L1-L2	Ethereum	0.0004 units	ϵ	ETH&ERC20	yes	easy
	BSC	0.000075 units	ϵ	BNB&BEP-20	yes	easy
	Tron	0.027 units	ϵ	Tron&TRC20	yes	easy

^a For the public mainnet, data is fetched from the respective blockchain explorer.^b ϵ can be either zero or close to zero, depending on the protocol specifications.^c BSC refers to the Binance Smart Chain.

3.2.1. Uncharged approaches for decentralized AI services

When trying to design an uncharged platform for decentralized AI services like the "free" YouTube or Gmail in traditional internet, we need to keep in mind that there is no such thing as a free lunch. So, who is actually paying for the AI services and bandwidth provided by the network miners? Existing decentralized solutions all rely on one-time or monthly micropayments, creating transactional friction that discourages adoption. In practice, we typically see strong consumer resistance to micropayments in favor of no fees, flat fees, or one-time payments [4]. Therefore, to build such approaches, we need to solve the problem of guaranteeing "free" and high-quality services to users while ensuring that network miners are rewarded as they provide an increasing amount of services.

The approach we can take is to draw inspiration from the inflation model of the EOS storage design [12]. In this model, there is a certain percentage of annual inflation on the total coin/token supply of the ecosystem to ensure that miners get paid. Meanwhile the clients will need to lock the platform related coin/token into smart contracts in order to gain allowance of job requests. Service providers⁶ collectively provide the computational power and AI service capacity to those requests. For users to access AI services, they must stake their tokens in the smart contract designated for AI services. Think of this staking process as making a fully refundable security deposit. Users can retrieve their tokens by releasing the service providers from the obligation to provide further AI services to them. This mechanism of staking/locking tokens from the client side will prevent all forms of Sybil attacks, which could flood the system with unlimited requests, halting the system indefinitely. Clients can only secure more service capacity by pledging more tokens to the network compared to other clients.

The system's robustness in general There are several major aspects to consider when designing a staking-and-use mechanism like this. First, we must ensure that the miners are motivated to provide honest and high-quality services to the clients. Not only should they possess adequate facilities and resources, such as substantial computational power and network bandwidth, but they should also have associated reputation records. These records would include scores given by clients for their services and the amount of stake they have locked, representing their commitment to the system's overall ecosystem. The system should also keep a record of clients' reviews. If a client continuously gives malicious reviews that deviate significantly from other clients' feedback, the system should implement appropriate penalty mechanisms for such behavior.

The miner's perspectives When a client sends a request, the system forwards it to a specific service provider. Assuming the service provider is designed to handle many requests simultaneously, there may be times when the number of incoming requests exceeds its pro-

cessing capacity. In such instances, requests are queued. The service provider would then prioritize these requests based on the number of tokens each client has staked. We consider that using a weighted round-robin (WRR) scheduling [2] to ensure more predictable and fair access, while still respecting the proportional stake.

Let's consider that at a specific time t , we have a miner k receiving requests from N_t clients, where each client sends a specific number of requests denoted as $(n_1^k, n_2^k, \dots, n_{N_t}^k)$. Each client has staked tokens in the amounts $(s_1, s_2, \dots, s_{N_t})$. We can determine the corresponding weight of each client using:

$$w_r = \lfloor \frac{s_r}{s_{\min}} \rfloor \quad (1)$$

where $s_{\min} = \min(s_1, s_2, \dots, s_{N_t})$

Once the weight is determined, miner k will stop accepting further requests by setting the `status` variable to `busy`. This signals the coordinator nodes to stop forwarding more requests to miner k . Let $w_{\max} = \max(w_1, w_2, \dots, w_{N_t})$. With interleaved WRR, miner k would require w_{\max} rounds to process all of the requests. In each round, one request from each client is processed. Assume there are N_1 clients with only one request. Then, in the first round, there are N_t requests to be processed, and in the second round, there are $N_t - N_1$ requests. This procedure continues until all requests have been iteratively processed. Afterward, the `status` variable of miner k is reset to `ready`.

Once a job request is processed, the global ledger receives a signed message from the service provider indicating successful completion, and the client obtains the AI service output from the miner. The client can then review the service provided by the miner by sending a signed message to the global ledger that reflects the quality of the service. While proof-of-reputation (POR) is generally used in existing literature as a method to produce blocks and validate transactions [13], we employ the reputation system as a reference for both the global ledger and clients. This may be a desired input for certain utility functions and protocols. Given ratings Good (G): 1, Fair (F): 0, and Bad (B): -1, the cumulative reputation for miner (service provider). $R_k(t)$ is computed as:

$$R_k(t) = 100 \times \frac{1}{1 + e^{-\theta \sum_{i=1}^N c_i}} \quad (2)$$

where c_i represents the latest reputation score given by client i and θ adjusts the sensitivity of the score. This logistic function maps any real number to the range [0, 100], ensuring a bounded and smooth reputation score.

Rewards for miners As we have discussed the inflation model, we need to dive into how these rewards can be distributed to individual miners. In traditional inflation models, such as the Solana ecosystem [30], rewards are proportionally distributed to validators based on

⁶Service providers and miners can be used interchangeably in this paper's context.

524 the volume of their staked tokens. In our case, however, the reward
 525 system is slightly more complex. Miners are rewarded not only for
 526 their computation but also, and more significantly, for the quality
 527 of the services they provide. This is determined by their reputation,
 528 service provision logs, and corresponding work volume calculated by
 529 the global ledger. We suggest that miners be rewarded based on their
 530 contribution, denoted as C_k for the miner indexed k , during a certain
 531 time span between t_1 and t_2 , as calculated by the following formula:

$$C_k = \sum_{t=t_1}^{t_2} \sum_{j=1}^M N_{\text{processed},t,j} \times W_{\text{service},t,j} \quad (3)$$

532 where $N_{\text{processed},t,j}$ denotes the number of requests processed for the
 533 j^{th} service at a specific time t . t_1 corresponds to the time of the last
 534 reward distribution, and t_2 represents the time of the next reward dis-
 535 tribution. Meanwhile, $W_{\text{service},t,j}$ signifies the weight associated with
 536 the j^{th} service's processed requests at that same time. For instance,
 537 the output of text-related services generally has a lower weight than
 538 that of image-related services due to its computational and memory
 539 requirements. Across the entire interval, we're considering M distinct
 540 services in the marketplace, while any new services can be added
 541 with weights determined by the community DAO.

542 **The client's perspectives** The system enhances user-friendliness on
 543 the client side by minimizing the necessary work and associated fees.
 544 Typically, clients can employ built-in third-party tools or APIs, such
 545 as Metamask and TrustWallet, to initiate processes. By staking any
 546 cryptocurrency recognized by the L1-L2 network, clients gain access
 547 to all AI services available on the platform with just a few clicks. The
 548 only cost is the initial staking transaction fee on L2, which can be
 549 less than one dollar on public chains like BSC or Polygon. Clients
 550 also have the option to rate the services upon receiving content from
 551 service providers, submitting their ratings to the coordinator nodes.
 552 Furthermore, to prevent DDoS attacks, coordinator nodes can offer
 553 the service without any charges but may set a request threshold for
 554 each client.

555 The platform also makes it possible to stake any crypto assets by
 556 leveraging the L1-L2 infrastructure. This means that as long as the L2
 557 is constructed on any of the mainstream public mainnets, individuals
 558 can stake not only bitcoin (wrapped BTC) [19], Ethereum/BNB, and
 559 stable coins such as USDT and USDC but also a variety of other assets
 560 to access services. However, there's a difference in the value of staked
 561 assets and the 'bandwidth' of services one can earn when using the
 562 native AI tokens compared to other cryptocurrencies. This difference
 563 is defined by a q ratio. Typically, q is valued at 0.1. So, for every one
 564 dollar's worth of AI tokens and other cryptocurrencies staked, the
 565 service volume ratio stands at 10:1. Both service providers and the
 566 global ledger adhere to this ratio. The rationale behind this design
 567 is to encourage more users to adopt the native tokens, promoting its
 568 market utility and commercialization.

569 **The coordinator's aspect and system securities** Coordinator nodes
 570 are responsible for directing messages and managing scheduling
 571 across the network, bridging the communication between clients and
 572 miners. While clients and miners interact seemingly directly, their
 573 exchanges are actually scheduled by these coordinators. Additionally,
 574 during each reward cycle, the coordinator nodes distribute system
 575 rewards to miners based on their respective contributions. Suppose
 576 that R denotes the total rewards to be distributed in the given time
 577 span, C_k represents the contribution of the miner k , and E is the
 578 overall count of miners. The aggregate contribution from all miners
 579 is represented by C_{total} , calculated as $C_{\text{total}} = \sum_{k=1}^E C_k$. Based on
 580 these parameters, the reward allocated to each individual miner k ,
 581 expressed as R_k , is given by:

$$R_k = \frac{C_k}{C_{\text{total}}} \times R \quad (4)$$

582 Securities There are several security concerns with the current
 583 reward distribution protocol. Clients might unjustly give negative
 584 reviews to honest service providers. Malicious miners might either
 585 refuse to provide services or deliver incorrect service outputs, and it's
 586 impossible to prevent and verify this given the system's distributed
 587 nature. In a more sophisticated attack, an entity might act as both
 588 a miner and a client to exploit the reward protocol. For example,
 589 they could flood the system with fake service exchange messages,
 590 artificially inflating a miner's contribution. The attackers can then
 591 get away with a significant portion of the system's periodic rewards,
 592 discouraging genuine miners.

593 Regarding malicious clients, the system can compare malicious
 594 reviews with other feedback. If a particular client's reviews consist-
 595 ently diverge from the majority, for example, if others frequently
 596 rate a service as "GOOD" while this client rates it as "BAD" or vice
 597 versa, access to the reputation protocol might be restricted for a set
 598 duration $t_{\text{restricted}}$. If the client continues to submit biased reviews, the
 599 restriction period can increase exponentially: $2t_{\text{restricted}}$, $4t_{\text{restricted}}$, and
 600 so on, thus safeguarding the system against potential harm from the
 601 client's side.

602 Malicious miners can initiate denial of service attacks or provide
 603 low-quality services to a subset of clients. However, such actions will
 604 quickly lead to negative feedback on their reputation score. While
 605 reputation scores do not directly affect a miner's contribution and
 606 rewards, they can influence the likelihood of a miner being assigned a
 607 task. Since reputation scores can be publicly accessed from coordi-
 608 nator nodes, miners with poor reputations are less likely to be scheduled
 609 for a request or chosen by a client as a service provider.

610 Attacking the reward protocol involves an entity acting as both
 611 miners and clients, continually updating the system with false and
 612 non-existent service exchanges, maliciously building up its contribu-
 613 tion over time. To counter this, two preventative measures are
 614 recommended. First, instead of allowing clients to choose the service
 615 provider directly, the coordinator nodes will select service providers
 616 based on their status and reputation. Nodes with higher reputation
 617 scores are more likely to be chosen. The random selection mech-
 618 anism employed by the coordinator nodes can be inspired by the
 619 verifiable random function (VRF) from Chainlink [9]. To integrate a
 620 reputation-weighted random selection using VRF, one begins by gen-
 621 erating an unpredictable and verifiable random number using VRF.
 622 This number is then used as input for a weighted random selection
 623 algorithm, where each of the K miners has a selection weight deter-
 624 mined by its reputation score. Techniques like the roulette wheel or
 625 stochastic acceptance can be utilized, ensuring that nodes with higher
 626 reputation scores have a higher probability of being selected. We call
 627 this function the weighted verifiable random function (AVRF). If we
 628 denote the selected miner index I_{request} as the routed miner for the
 629 request, then the AVRF can be written as:

$$I_{\text{request}} = \psi(\text{VRF}, R) \quad (5)$$

630 where $R = (r_1, r_2, \dots, r_K)$ represent the reputation scores of the K
 631 miners in the system, with r_i denoting the reputation score of the
 632 i^{th} miner. Once clients are unaware of which miners might serve
 633 their requests, they lack the motivation to initiate the reward protocol
 634 attacks, as they might inadvertently boost the contributions of other
 635 miners. Moreover, the system can impose a threshold on the number
 636 of requests each client can make for specific AI services to prevent
 637 system flooding. Once this threshold is reached, the client will be
 638 temporarily frozen before it can send another request.

639 **Economics** With the Uncharged Protocol, all token holders will be
 640 contributing to this system via a portion of the 5-10% annual token
 641 inflation. Specifically, those who wish to access services must lock
 642 up their tokens, rendering them unable to sell these tokens until they
 643 finish using the service. Clients requiring long-term or continuous
 644 services may lock up their tokens for an indefinite period.

645 As the demand for services increases, leading to more tokens being

locked up at a rate higher than the inflation rate due to the platform's growing market and commercial scale, the token economy undergoes an effective monetary deflation. This deflationary trend increases the value of tokens earned by service providers, encouraging them to offer a broader range of superior services.

Should there be a substantial decrease in service demand, the released tokens might flood the market, leading to an effective price drop beyond the standard inflation rate. This means the value of tokens may decrease, and the quality or quantity of services that providers can afford to offer might decline. However, due to the reduced demand, providers could choose to scale down their service offerings, thus reducing operational costs. Alternatively, adjustments could be made to the staking mechanism, recalibrating the number of tokens a client needs to stake to access a service.

Ultimately, clients in need of services fund the ecosystem via the time-value of their staked tokens. This ensures a smooth user experience with no micropayments, no transactional hurdles, and no unexpected fees.

3.2.2. Charged approaches for decentralized AI services

Compared to uncharged approaches, charged approaches are more straightforward. The major process is that clients compensate service providers through subscription fees. These can be achieved per service request or as a single payment covering monthly or yearly durations. This approach is closely tied to the specific service being provided, as clients usually pay for a singular type of service to access. Although this design is simpler at the system level, with fewer security and economic complexities, there's an important responsibility for the platform: guiding clients to reputable service providers. Some providers might maliciously take clients' funds and later come back online with a new identity to commit fraud repeatedly. Occasional outages, even if unintended, can damage user experiences and decrease trust in the platform. As a result, it's crucial for the system to showcase trustworthy service providers prominently. Although some users might have specific preferences, the system should always highlight reliable service sources. Table 2 outlines the factors the platform considers when listing service providers, with varying importance depending on the situation.

Table 2. Factors in Service Provider Display

Factors	Data Structure Types
Customer Rating of Service	float (typically 1-5/1-10)
Total Number of Subscribers	integer (0 to max clients)
Total Number of Tokens Staked	float value s
Types of AI Services Provided	string list T

Customer rating of service is typically represented as a float, indicating a one-star to five-star rating. The types of AI services provided include a string list that indicates the AI services the service provider supports. Like the uncharged approaches, a reputational protocol is also involved. Users are required to rate their service providers by submitting signed ratings to the global ledger. The global ledger then updates the miner's reputation based on these new ratings. Additionally, the coordinator node receives a portion of the payment as transaction fees. These fees typically cover the coordinator's operational costs and rewards but are kept moderate to prevent transaction friction and discourage miners. This fee structure also helps prevent the reward cheating attack mentioned earlier.

3.3. Protocol implementations: network participants and process flow

We focus on the operations carried out by various participants: clients, network coordinators, and miners. We illustrate the process flow of different algorithms. To ensure the protocol accommodates as many types of AI services as possible, we have integrated both charged and uncharged economic approaches. Reputation protocols include

Algorithm 1 Uncharged Protocol for Decentralized AI Services

```

1: function ACQUIRESERVICEPASS(client_token)
2:   if LockTokensIntoSmartContract(client_token) then
3:     return GenerateServicePass()
4:   else
5:     return "Insufficient stake or token lock failed"
6:   end if
7: end function
8: procedure REQUESTSERVICE(service_pass, request)
9:   if IsValidServicePass(service_pass) then
10:    provider ← SelectServiceProvider()
11:    if provider.status = ready then
12:      service_output ← provider.Serve(request)
13:      feedback ← GetClientFeedback(service_output)
14:      REVIEWSERVICE(provider.id, feedback)
15:    else
16:      QueueRequest(request)
17:    end if
18:  else
19:    return "Invalid Service Pass"
20:  end if
21: end procedure
22: function SELECTSERVICEPROVIDER
23:   vrf_value ← GenerateVRF()
24:   provider_index ← AVRF(vrf_value, R)
25:   return provider_list[provider_index]
26: end function
27: function GETCLIENTFEEDBACK(service_output)
28:   return ClientFeedback(service_output)
29: end function
30: procedure REVIEWSERVICE(provider_id, feedback)
31:   ledger.UpdateReputation(provider_id, feedback)
32: end procedure

```

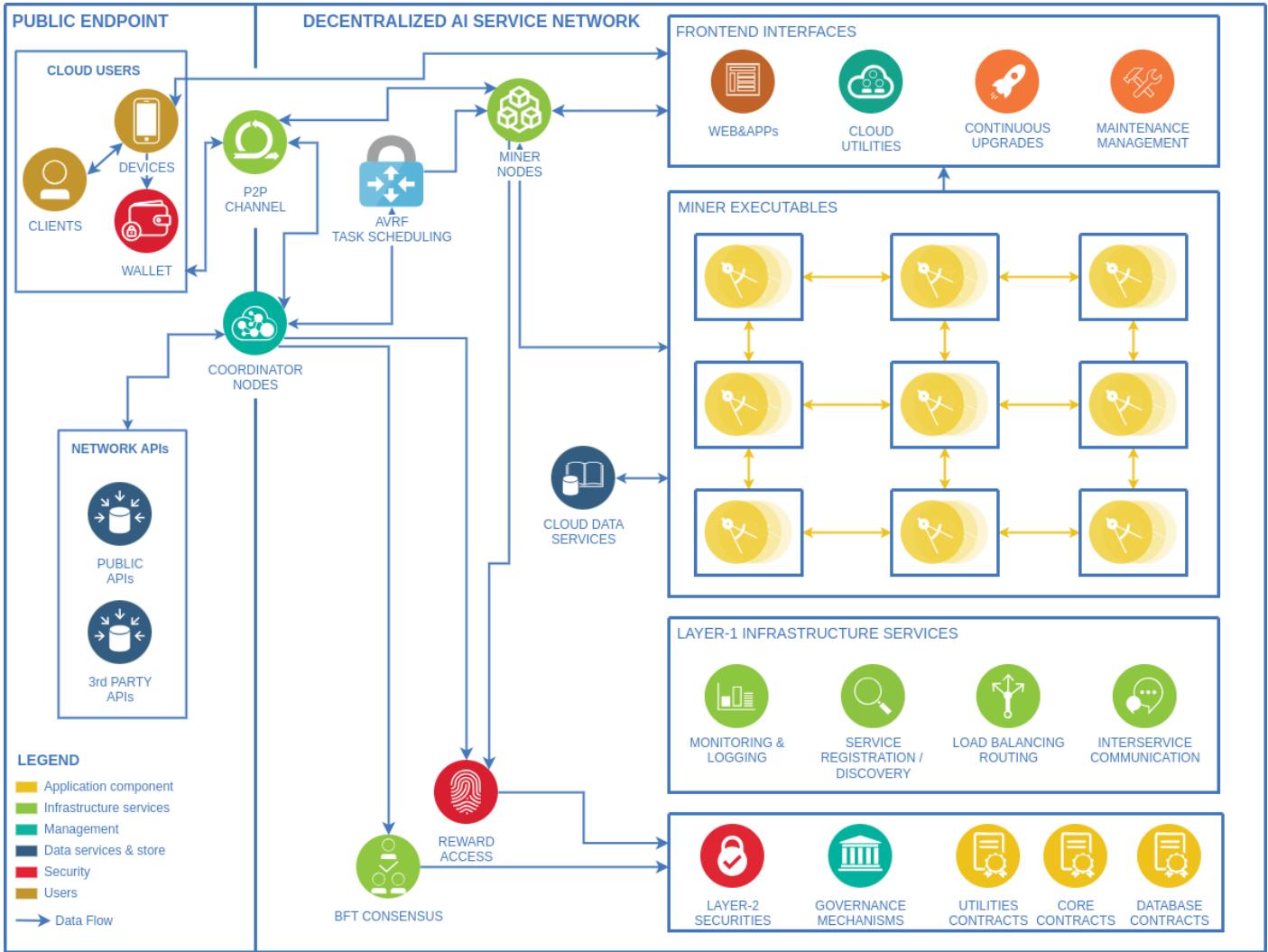


Figure 1. Overview of the Decentralized AI Service Network: This schematic represents the interconnected components of a decentralized AI service system. **On the left**, we have the public endpoint catering to users and devices, managed through wallets and coordinated via P2P channels and task scheduling. There are also network APIs for the convenience of developers and product development. **On the right**, frontend interfaces serve as the communication bridge between clients and the network, enabling easier integration for the community based on decentralized AI functionalities. The core of service in the network includes miner nodes and executables, which perform computations and return results. This entire ecosystem is supported by various infrastructure services, such as cloud data services, L1-L2 infrastructure and APIs. Layer-1 focuses on primary infrastructure services, with distinct utilities, while Layer-2 emphasizes securities and governance. All components work interactively and cooperatively, ensuring efficient data flow, task distribution, and reward mechanisms.

rating scores from both sets of clients. This protocol is designed to enable large-scale commercialization.

The global ledger, coordinator nodes and peer to peer connections
 In our decentralized AI service network, the Global Ledger \mathcal{L} plays a key role as a system record, logging all essential network interactions. The ledger contains three key components: the *orders record*, the *task cycle data*, and the *node info*. The *orders record* logs all orders placed by clients within the network, each containing the specific task details requested by a client; this includes the required service, data, and associated rewards in the charged case. The *task cycle data* records the metadata of tasks that have undergone the full cycle of AI service provision and reward distribution in the network; it includes service generation signatures, related value exchanges, and potential comments and ratings. The *node info* section saves the details of all registered service providers within the network, including their reputation and performance history. Collectively, these components of the ledger boost the network's performance by ensuring all operations are traceable and accessible in a timely manner. The *coordinator node*, with the responsibility of publishing multi-signature transactions on the blockchain and updating contract states, plays a central role in managing ledger data and global states. Through the application

of Byzantine Fault Tolerance (BFT) consensus, such as the Practical Byzantine Fault Tolerance (PBFT) algorithm [3], it effectively maintains, updates, and synchronizes the Global Ledger \mathcal{L} . Besides storing and managing a synchronized copy of the global ledger, the coordinator nodes also act as data access points for other network participants. They provide on-demand access to the global ledger, ensuring its data is always available for different network operations.

Miners in the network are responsible for providing reliable data transfer links to supply AI service outputs. This data must be consistently accessible throughout the task cycle. Failure to do so can lead to filled orders receiving negative comments. It is also the client's responsibility to download the necessary data to their local storage for efficient service exchange processes.

Client cycle We provide an overview of the client cycle, which primarily includes the Put, Get, and Rate protocols. This offers a complete cycle from initiating the request to receiving a response and subsequently commenting on the services.

1. **Put:** *The client put orders requesting AI service.* Clients can request AI services from the network using staked pass or utility tokens. A client initiates the Put process by submitting an order to the network. Subsequently, coordinators have the authority

744 to decide which service provider will handle the order, unless it's
 745 pre-specified in the charged case. Coordinator nodes submit job
 746 allocation messages to the global ledger. Clients can choose be-
 747 tween free or charged services by providing relevant information
 748 in the appropriate sections of the order message. The selection
 749 of a paid service might result in higher quality service outputs.

750 **2. Get:** *Client retrieves model from the network.* Clients can access
 751 the AI service output from the network as soon as it's complete.
 752 A direct peer-to-peer link connects the client and miner, initiated
 753 when the job allocation is logged in the network by coordinator
 754 nodes. The service provider then sends the AI results directly to
 755 the client. Once done, a confirmation is sent to the coordinator,
 756 which updates the global ledger with a completed job task. It is
 757 the miners' responsibility to ensure that their outputs are always
 758 made available to the clients to avoid negative effects on their
 759 reputations in the network.
 760 **3. Rate:** Ratings are important components in the network, provid-
 761 ing global ledger with the service quality feedback of different
 762 service providers. Clients can initiate **Rate** by sending a signed
 763 message to the global ledger commenting on the latest service
 764 output of a service provider. *Note:* the reputation score of a
 765 certain client on a service provider is always based on the lat-
 766 est ratings, meaning their previous comments are deleted and
 767 updated with the latest one.

768 **Mining Cycle (for service providers)** We give an overview of the
 769 mining cycle of service providers. Service providers earn rewards by
 770 competing to earn higher reputation score in the network.

771 **1. Register:** Service Providers pledge their computational re-
 772 sources to the network. This is done by depositing collateral, via
 773 a transaction in the network, using **Miner.RegisterResource**.
 774 This collateral is locked in for the time intended to provide the
 775 service, and is returned upon request of the service provider if
 776 the provider decides to stop committing to the network, using
 777 **Miner.UnRegisterResource**. Once the service provider is reg-
 778 istered, they can start generating model claims which will be
 779 added to the global ledger.

780 **Miner.RegisterResource/UnRegisterResource**
 781 • INPUTS:
 782 – current global ledger \mathcal{L}_t
 783 – registration request **register**
 784 • OUTPUTS: current global ledger \mathcal{L}'_t

785 **2. Service Executables:** After registration, the service providers
 786 execute various AI services offered by the network infrastructure,
 787 primarily determined by their capability to manage different
 788 services. The greater their computing power and bandwidth,
 789 the more services they can offer. The network schedules the
 790 distribution of task requests, and miners can accept incoming
 791 tasks once their service is online. After generating the output,
 792 the miner sends it back to the client and notifies the network
 793 that the specific request has been handled, thereby claiming
 794 their contribution.

795 **Miner.Exec**
 796 • INPUTS:
 797 – current orders from global ledger \mathcal{L}_t
 798 • OUTPUTS: signed message **sClaim**

799 **3. Sending Outputs:** Service Providers are responsible for en-
 800 suring the availability of the generated output for a client g_c
 801 throughout the full mining cycle. This is done through the
 802 **Miner.SendOutput** function. If a service provider fails to main-
 803 tain the availability of these data, the network may invalidate the
 804 service, which will result in the service provider not receiving
 805 the contribution rewards. However, if a miner claims a contribu-
 806 tion but doesn't provide the output to the client, it may lead to
 807 negative reviews, thus affecting the miner's operation of services.

808 **Miner.SendOutput**
 809 • INPUTS:
 810 – order ID **oID**
 811 – generated output g_c
 812 • OUTPUTS: success status **sStatus**

813 **3.4. Discussions**

814 **3.4.1. Capability and scalability**

815 The system's transaction throughput performance, or in other words,
 816 the amount of information the system can process per second, is
 817 determined by its underlying design structure. The coordinator nodes
 818 act as a platform within the system, coordinating between clients and
 819 miners and enabling self-governance to initiate, process, and finalize
 820 services. Although the actual influx of transactions (including orders,
 821 confirmations, and ratings) will largely depend on the customer base
 822 and the total hash power of the network, the processing capability of
 823 the global states maintained by the coordinator nodes can be analyzed.
 824 Given the sizes of the orders, confirmations, and ratings messages, it
 825 can be estimated that a global ledger maintained by 50 coordinator
 826 nodes distributed worldwide can synchronize approximately 1000
 827 full task cycles per second, as evidenced by the real L1-L2 network
 828 results in [39].

829 A significant advantage of the design lies in the allocation of
 830 computation-intensive tasks and storage to network participants.
 831 This strategy avoids overconsuming global storage and bandwidth,
 832 which could become costly, especially since updating global states
 833 is a synchronous process. The global ledger only stores information
 834 about orders, confirmations, and ratings, each of which is measured
 835 in kilobytes. Moreover, processing this information requires a com-
 836 putational complexity of $\mathcal{O}(1)$. Such a design enables the system
 837 to handle a virtually limitless number of task requests and service
 838 finalizations concurrently.

839 **3.4.2. Security**

840 In most web3 protocols, the security of a protocol is guaranteed by
 841 economic incentives, i.e., attacking the system is more costly than
 842 complying with it. Similarly, in the designed platform, one would
 843 need to obtain more tokens than the counterparties to initiate attacks,
 844 which can often prove quite expensive. Unless the potential rewards
 845 are substantial, there is little incentive for someone to attack the
 846 protocol.

847 In a scenario where the attack comes from the coordinator's side,
 848 it involves tampering with the rewarding process in the coordinator
 849 nodes' global ledger. This allows hackers to withdraw all tokens
 850 from the rewards distribution contract. To compromise the multi-sig
 851 design of the L1-L2 infrastructure, the attackers would need a (m/c)
 852 portion of the total staked tokens by the coordinator nodes. We call
 853 this *Linear staking impact*, meaning that to be successful, an attacker
 854 must have a budget B greater than a (m/c) portion of the combined
 855 staked tokens of all coordinator nodes. More precisely, we mean
 856 that as a function of m , $B(m) = dm$ in a network of c coordinator
 857 nodes, each with a fixed staked amount d . Given our requirement for
 858 coordinator nodes to stake a significant amount of tokens to act as
 859 network coordinators, a hacker would need at least 10% of the total
 860 circulation if 20% of tokens are held by the honest coordinator nodes
 861 (assuming $m = 18$ and $c = 30$). Therefore, the cost of such an attack
 862 is generally much higher than the tokens in the reward contract.

863 In a scenario where the attack originates from the client's side,
 864 it involves flooding the system with requests using multiple fake
 865 identities, thereby halting the system by consuming all its resources.
 866 As discussed in the design sections, this is prevented by WRR, where
 867 a fake identity will be served only once or twice in a time frame while
 868 many others are being served. To effectively flood the system, they
 869 would need to increase their tokens, incurring a much higher cost.
 870 Another common attack pattern involves giving malicious reviews to
 871 honest miners during a service. However, if a client's ratings deviate
 872 significantly from the majority of reviews most of the time, the client

862 might be suspended from commenting for a period of time by the
 863 coordinator nodes to mitigate its negative influence.

864 In scenarios where the attacker originates from the miner's side,
 865 it typically involves miners attempting to maliciously increase their
 866 contribution, as this directly correlates to the distribution of rewards
 867 for uncharged services. This contribution can be built up through two
 868 variables: the service weight (determined directly by the community
 869 DAO) and the number of services provided by the miner. While it
 870 is extremely difficult to manipulate the service weight, miners may
 871 be incentivized to exaggerate the number of services they provide.
 872 Fake clients might artificially inflate the number of services, but this
 873 is countered by the AVRF scheduler and a threshold for the number
 874 of service accesses in a given timeframe. Moreover, fake requests
 875 will inadvertently boost the contributions of others in the current
 876 protocol, thereby offsetting the negative impact. A miner might also
 877 try to execute a denial of service attack, but their reputation would
 878 rapidly decline due to subpar service quality. Additionally, commu-
 879 nity members can vote out malicious miners via the DAO. But most
 880 importantly, platform developers should design an intricate recom-
 881 mendation and list algorithm that prioritizes statistically reliable and
 882 honest service providers on the frontend (be it a website or app). This
 883 ensures that users are more likely to select top-tier service providers,
 884 as these are presented with priority. As a result, if these prioritized
 885 providers offer charged services, they are more likely to be trusted
 886 by clients to be committed to the service deal, rather than "rugging"
 887 once they receive payment.

888 In general, in a staking-intensive web3 environment, many attack
 889 types can be mitigated by adjusting various staking protocols and the
 890 time required for the staking and unstaking processes. To the best of
 891 our knowledge, the system is robust against different types of attacks.
 892 The fundamental idea is that we aim to keep the cost of attacking
 893 the system high at all times, regardless of the angle from which the
 894 attack may originate.

895 **3.4.3. Advantages**

896 We believe one of the major advantages of the solution lies in its
 897 consensus mechanism design. This provides significant capacity and
 898 scalability benefits compared to other solutions in this field. Gener-
 899 ally, Web3 infrastructures are less efficient due to their distributed
 900 nature and inherent lack of trust. However, with this design, the
 901 network coordinator, which maintains the global ledger and global
 902 states, is relieved from handling large data storage or the heavy com-
 903 putation tasks common in most AI servicing processes. Instead, these
 904 tasks are delegated to participant nodes with ample resources. Par-
 905 ticipants are given strong cryptoeconomic incentives to act honestly
 906 and diligently. This creates a system that is largely self-governing,
 907 further enhancing the solution's capacity and scalability. Beyond the
 908 reputation protocol, participants are regulated by a community DAO.
 909 For instance, if any service provider fails to provide the necessary
 910 service and bandwidth for a smooth exchange process, they may face
 911 penalties. This could come in the form of other nodes on the net-
 912 work voting against them in a community motion within the DAO.
 913 Consequently, the consensus mechanism ensures that participants
 914 remain committed to their orders and services, guaranteeing system
 915 liveliness.

916 Another major advantage of the solution, compared to others, is
 917 the design of its L1-L2 system structure, which ensures easy system
 918 upgradability. AI is a rapidly changing industry, with new types of
 919 services emerging daily. The protocol employs Layer-2 (on-chain) ap-
 920 plications for depositing, withdrawing, and transferring users' assets,
 921 while the majority of operations are conducted on Layer-1 (off-chain)
 922 to facilitate upgradability. To integrate new services, miners and co-
 923 ordinators can simply upgrade to the latest version of the software.
 924 Subsequently, clients will have the ability to specify these new service
 925 types in their orders. In theory, the system can incorporate any kind
 926 of AI service into the L1 infrastructure.

4. Conclusions and Future Works

927 In this work, we have provided a comprehensive review of the op-
 928 portunities and challenges related to merging web3 and the AI mar-
 929 ketplace. We thoroughly studied the advantages of both fields and
 930 the challenges involved in their integration. We also presented our
 931 solutions on this topic, with a primary focus on the framework's
 932 commercial rationality, security, and efficiency. We believe that the
 933 framework should first demonstrate feasibility and the potential to
 934 catalyze large-scale commercialization before focusing on its security
 935 and efficiency. We began with the user experience in mind and then
 936 identified ways to technically realize our vision. In general, we've
 937 made contributions in two main areas: firstly, we offer an overview of
 938 the commercial landscape of web3 and the AI marketplace, highlight-
 939 ing both opportunities and challenges in this commercial avenue,
 940 and secondly, we proposed our solutions.

941 To our knowledge, our platform, which supports both charged and
 942 uncharged AI services, is the first in the industry to introduce such
 943 a framework with the key protocols presented. It emphasizes user
 944 experience, maximizing its potential for widespread adoption, yet it
 945 is intricately designed to ensure the platform's resilience against the
 946 various types of attacks prevalent in the web3 industry. While the
 947 web3 ecosystem is occasionally perceived as less efficient by the web2
 948 community, our system's total throughput demonstrates potential in
 949 serving customers worldwide. We showed that with a optimized de-
 950 sign structure, high-efficiency web3 platforms can be realized without
 951 compromising their distributed nature, thus ensuring broader acces-
 952 sibility. In summary, we have proposed a solution that allows anyone
 953 holding cryptocurrencies to access a range of AI services, whether
 954 they seek free offerings or wish to pay for customized services.

955 One aspect not covered in this paper is the execution of experiments
 956 involving different sub-protocols within the designed framework, es-
 957 pecially those concerning the interaction between clients and miners
 958 as actual tasks are resolved. This omission is primarily because any
 959 simulation in this regard would merely represent a specific case of the
 960 system's capacity and throughput. However, analyzing the protocol
 961 from a financial perspective is set as part of our future work. We
 962 aim to engage the current crypto mining infrastructure in the web3
 963 community by introducing network utility tokens and implementing
 964 a comprehensive version of the framework with detailed parameters.
 965 This would allow for a thorough analysis of the system's performance
 966 on real-world tasks, paving the way for further developments and
 967 deeper understanding of the AI marketplace within web3.

968 **5. Contact us**

969 You can contact us through these methods.

970  [MintAI X](#)
 971  develop@mintai.network

975 **6. Supporting**

976 Did you like this topic? Check out our latest project named [MintAI](#)
 977 [Network](#), aiming to build the largest AI aggregator built on web3!

978 **Any contributions are welcome!**

979 If you wish to support my work, you can do so through contributing
 980 to our reference implementation of the MintAI protocol:
 981 <https://github.com/DeAI-Artist/MintAI>.

982 **References**

983 [1] M. J. Fischer, N. A. Lynch, and M. S. Paterson, "Impossibility of
 984 distributed consensus with one faulty process", *J. ACM*, vol. 32,
 985 no. 2, pp. 374–382, Apr. 1985, ISSN: 0004-5411. DOI: [10.1145/3149.214121](https://doi.org/10.1145/3149.214121). [Online]. Available: <https://doi.org/10.1145/3149.214121>.

[2] M. Katevenis, S. Sidiropoulos, and C. Courcoubetis, "Weighted round-robin cell multiplexing in a general-purpose atm switch chip", *IEEE Journal on Selected Areas in Communications*, vol. 9, no. 8, pp. 1265–1279, 1991. DOI: [10.1109/49.105173](https://doi.org/10.1109/49.105173).

[3] M. Castro and B. Liskov, "Practical byzantine fault tolerance", *OSDI*, vol. 99, pp. 173–186, 1999.

[4] B. Kahin and H. R. Varian, "Fixed-fee versus unit pricing for information goods: Competition, equilibria, and price wars", in *Internet Publishing and Beyond: The Economics of Digital Information and Intellectual Property*. 2000, pp. 167–189.

[5] "Bitcoin: A peer-to-peer electronic cash system", *Cryptography Mailing list* at <https://metzdowd.com>, Mar. 2009. [Online]. Available: <https://www.bitcoin.org>.

[6] J. Benet, "IPFS - content addressed, versioned, P2P file system", *CoRR*, vol. abs/1407.3561, 2014. arXiv: [1407.3561](https://arxiv.org/abs/1407.3561). [Online]. Available: [http://arxiv.org/abs/1407.3561](https://arxiv.org/abs/1407.3561).

[7] D. Schwartz, N. Youngs, A. Britto, *et al.*, "The ripple protocol consensus algorithm", *Ripple Labs Inc White Paper*, vol. 5, no. 8, p. 151, 2014.

[8] D. Silver, A. Huang, C. Maddison, *et al.*, "Mastering the game of go with deep neural networks and tree search", *Nature*, vol. 529, pp. 484–489, 2016, Received: 11 November 2015; Accepted: 05 January 2016; Published: 27 January 2016; Issue Date: 28 January 2016. DOI: [10.1038/nature16961](https://doi.org/10.1038/nature16961). [Online]. Available: <https://doi.org/10.1038/nature16961>.

[9] S. Ellis, A. Juels, and S. Nazarov. "Chainlink: A decentralized oracle network". version v1.0. Accessed: September 27, 2023. (2017), [Online]. Available: <https://research.chain.link/whitepaper-v1.pdf>.

[10] P. Labs, "Filecoin: A decentralized storage network", Protocol Labs, Whitepaper, Jul. 2017.

[11] L. M. Bach, B. Mihaljevic, and M. Zagar, "Comparative analysis of blockchain consensus algorithms", in *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 2018, pp. 1545–1550. DOI: [10.23919/MIPRO.2018.8400278](https://doi.org/10.23919/MIPRO.2018.8400278).

[12] EOS.IO. "Eos.io technical white paper v2". (Mar. 2018), [Online]. Available: <https://github.com/EOSIO/Documentation/blob/master/TechnicalWhitePaper.md>.

[13] F. Gai, B. Wang, W. Deng, and W. Peng, "Proof of reputation: A reputation-based consensus protocol for peer-to-peer network", in *Database Systems for Advanced Applications*, J. Pei, Y. Manolopoulos, S. Sadiq, and J. Li, Eds., Cham: Springer International Publishing, 2018, pp. 666–681.

[14] T. Nguyen and K. Kim, "A survey about consensus algorithms used in blockchain", *Journal of Information Processing Systems*, vol. 14, pp. 101–128, Jan. 2018. DOI: [10.3745/JIPS.01.0024](https://doi.org/10.3745/JIPS.01.0024).

[15] *Ocean Protocol: A Decentralized Substrate for AI Data and Services*, Accessed: September 27, 2023, 2018. [Online]. Available: <https://oceanprotocol.com/tech-whitepaper.pdf>.

[16] TRON DAO. "Advanced decentralized blockchain platform, Whitepaper version: 2.0, tron protocol version: 3.2". Accessed: [insert date you accessed the document], TRON. (Dec. 10, 2018), [Online]. Available: https://tron.network/static/doc/white_paper_v_2_0.pdf.

[17] A. Baldominos and Y. Saez, "Coin.ai: A proof-of-useful-work scheme for blockchain-based distributed deep learning", *Entropy*, vol. 21, no. 8, 2019, ISSN: 1099-4300. DOI: [10.3390/e21080723](https://doi.org/10.3390/e21080723). [Online]. Available: <https://www.mdpi.com/1099-4300/21/8/723>.

[18] *Fetch.AI Whitepaper*, Accessed: September 27, 2023, 2019. [Online]. Available: <https://whitepaper.io/document/447/fetch-whitepaper>.

[19] K. Network, B. Inc, and R. Protocol, "Wrapped tokens: A multi-institutional framework for tokenizing any asset", Whitepaper, version 0.2, Jan. 2019.

[20] W. Wang, H. Dinh Thai, P. Hu, *et al.*, "A survey on consensus mechanisms and mining strategy management in blockchain networks", *IEEE Access*, vol. PP, pp. 1–1, Jan. 2019. DOI: [10.1109/ACCESS.2019.2896108](https://doi.org/10.1109/ACCESS.2019.2896108).

[21] H. Al-Breiki, M. H. U. Rehman, K. Salah, and D. Svetinovic, "Trustworthy blockchain oracles: Review, comparison, and open research challenges", *IEEE Access*, vol. 8, pp. 85 675–85 685, 2020. DOI: [10.1109/ACCESS.2020.2992698](https://doi.org/10.1109/ACCESS.2020.2992698).

[22] D. Salah, M. H. Ahmed, and K. ElDahshan, "Blockchain applications in human resources management: Opportunities and challenges", in *Proceedings of the 24th International Conference on Evaluation and Assessment in Software Engineering*, ser. EASE '20, Trondheim, Norway: Association for Computing Machinery, 2020, pp. 383–389, ISBN: 9781450377317. DOI: [10.1145/3383219.3383274](https://doi.org/10.1145/3383219.3383274). [Online]. Available: <https://doi.org/10.1145/3383219.3383274>.

[23] T. W. Bank. "Credit card holders statistics". Accessed: 2023-09-27. (2021), [Online]. Available: <https://genderdata.worldbank.org/indicators/fin7-t-a/>.

[24] J. Boyle, *BTC Mining Used More Electricity than Sweden*, Accessed: September 27, 2023, 2021. [Online]. Available: <https://beincrypto.com/btc-mining-used-more-electricity-than-sweden/>.

[25] Y. Liu, Y. Lan, B. Li, C. Miao, and Z. Tian, "Proof of learning (pole): Empowering neural network training with consensus building on blockchains", *Computer Networks*, vol. 201, p. 108 594, 2021, ISSN: 1389-1286. DOI: <https://doi.org/10.1016/j.comnet.2021.108594>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1389128621004965>.

[26] Neo C K Yiu, "An overview of forks and coordination in blockchain development", 2021. DOI: [10.13140/RG.2.2.36579.07207](https://doi.org/10.13140/RG.2.2.36579.07207). [Online]. Available: [http://rgdoi.net/10.13140/RG.2.2.36579.07207](https://doi.org/10.13140/RG.2.2.36579.07207).

[27] Swiss Government, *Federal Act on the Adaptation of Federal Law to Developments in Distributed Electronic Register Technology*, Accessed: September 27, 2023, 2021. [Online]. Available: <https://www.aramis.admin.ch/Default?DocumentID=68053%5C&Load=true>.

[28] M. Salehi, J. Clark, and M. Mannan, *Not so immutable: Upgradability of smart contracts on ethereum*, 2022. arXiv: [2206.00716](https://arxiv.org/abs/2206.00716) [cs.CR].

[29] K. W. Win. "The state of decentralized storage". Accessed: September 27, 2023. (Oct. 2022), [Online]. Available: <https://www.coingecko.com/research/publications/the-state-of-decentralized-storage>.

[30] A. Yakovenko. "Solana: A new architecture for a high performance blockchain". version 0.8.13. Accessed: 2023-10-17. (2022), [Online]. Available: <https://solana.com/solana-whitepaper.pdf>.

[31] H. Alshahrani, N. Islam, D. Syed, *et al.*, "Sustainability in blockchain: A systematic literature review on scalability and power consumption issues", *Energies*, vol. 16, p. 1510, Feb. 2023. DOI: [10.3390/en16031510](https://doi.org/10.3390/en16031510).

1107 [32] “Artificial intelligence (ai) market (by offering: Hardware, soft-
 1108 ware, services; by technology: Machine learning, natural lan-
 1109 guage processing, context-aware computing, computer vision;
 1110 by deployment: On-premise, cloud; by organization size: Large
 1111 enterprises, small & medium enterprises; by business func-
 1112 tion: Marketing and sales, security, finance, law, human re-
 1113 source, other; by end-use:) - global industry analysis, size, share,
 1114 growth, trends, regional outlook, and forecast 2023-2032”, Pub-
 1115 lisher/Agency, Location if known, Tech. Rep., 2023.

1116 [33] Cointelegraph, *Cryptocurrency miners may lead the next stage*
 1117 of ai, Accessed: September 27, 2023, 2023. [Online]. Available:
 1118 <https://cointelegraph.com/news/cryptocurrency-miners-may-lead-the-next-stage-of-ai>.

1119 [34] M. Commerce, “Ai market by technology type, deployment
 1120 method, solution type, integration (technologies, networks,
 1121 and devices) and industry verticals 2023 - 2028”, Mind Com-
 1122 mmerce, Report, Feb. 2023, Region: Global, p. 248.

1123 [35] A. De Vries, “Cryptocurrencies on the road to sustainability:
 1124 Ethereum paving the way for bitcoin”, *Patterns*, vol. 4, no. 1,
 1125 p. 100 633, 2023, ISSN: 2666-3899. DOI: <https://doi.org/10.1016/j.patter.2022.100633>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2666389922002653>.

1126 [36] F. Duarte, “Number of chatgpt users (2023)”, *OpenAI News*,
 1127 Jul. 2023.

1128 [37] *Ethereum Network Hashrate Chart*, Accessed: September 27,
 1129 2023, 2023. [Online]. Available: <https://etherscan.io/chart/hashrate>.

1130 [38] Filfox, *Network overview*, English, Accessed: September 27,
 1131 2023, Mainnet, China: Filfox, 2023. [Online]. Available: <https://filfox.info/en>.

1132 [39] P. Li, *Proof of training (pot): Harnessing crypto mining power*
 1133 for distributed ai training, 2023. arXiv: 2307.07066 [cs.CR].

1134 [40] S. Noy and W. Zhang, “Experimental evidence on the productiv-
 1135 ity effects of generative artificial intelligence”, *Science*, vol. 381,
 1136 no. 6654, pp. 187–192, Jul. 2023. DOI: 10.1126/science.adh2586.
 1137 [Online]. Available: <https://science.sciencemag.org/content/381/6654/187>.

1138 [41] “Render network whitepaper”, Tech. Rep., May 2023, Origi-
 1139 nally posted August 28th, 2017.

1140 [42] *Singularitynet*, Accessed: 2023-04-05, 2023. [Online]. Available:
 1141 <https://singularitynet.io/>.

1142